(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－22751
（P2000－22751A）
(43)公開日　平成12年１月21日(2000.1.21)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｈ０４Ｌ　12/56		Ｈ０４Ｌ　11/20	１０２Ｆ	
Ｇ０６Ｆ　13/00	３５３	Ｇ０６Ｆ　13/00	３５３Ｃ	

審査請求　未請求　請求項の数28　ＯＬ　外国語出願　（全 27 頁）

(21)出願番号　特願平11－133112

(22)出願日　平成11年４月５日(1999.4.5)

(31)優先権主張番号　０９／０５５０３１

(32)優先日　平成10年４月３日(1998.4.3)

(33)優先権主張国　米国（ＵＳ）

(71)出願人　599065037
アルテオン　ネットワークス　インコーポレイテッド
アメリカ合衆国　カリフォルニア州　95119　サン　ホセ　サン　イグナチオ　アベニュー　6351

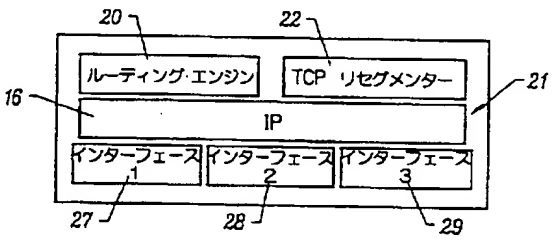
(72)発明者　セオドア　シュローダー
アメリカ合衆国　カリフォルニア州　95120　サン　ホセ　レンウッド　ウェイ　6961

(74)代理人　100059959
弁理士　中村　稔　（外６名）

最終頁に続く

(54)【発明の名称】　ＴＣＰリセグメンテーション

(57)【要約】　（修正有）
【課題】　ＴＣＰリセグメンテーションを提供する。
【解決手段】　リセグメンテーションエンティティがＴＣＰリセグメンテーションを実施し、受信ホストは、それがまるで受信ホストのＭＴＵに関して特に転送されるように見えるパケットを受信する。受信ホストは、ＩＰリアセンブリのためにＣＰＵ利用及びバッファリングを必要としない。送信ホストは、受信ステーションのＭＴＵに関係なく、その最大ＭＴＵでＴＣＰセグメントを転送し、中間のルーティングエンティティはＴＣＰリセグメンテーションが生じることを保証する。リセグメントされたＴＣＰセグメントを包含するＩＰデータグラムが失われる際、送信ホストは、損失し、完全なＴＣＰセグメントでないデータを転送する。



【特許請求の範囲】

【請求項１】電子ネットワークにわたる２又はそれ以上のホストの間で情報パケットを交換するための装置であって、少なくとも前記ホストのうちの１つが、前記２又はそれ以上のホストの他のものとサイズが異なる情報パケットを転送し且つ受信し、前記情報パケット内の大きな情報セグメントを、複数の対応する情報サブパケット内の複数の小さなサブセグメントにリセグメントするために、前記ホストの前記１又はそれ以上のものか、それらのそれぞれのネットワークインターフェースのいずれかに配置されたリセグメンテーションエンティティを有する、装置。

【請求項２】受信ホストは、情報パケットがまるでそれらが前記受信ホストに関して特に転送されるように見えるサイズのものである前記情報パケットを受信する、請求項１に記載の装置。

【請求項３】送信ホストが、情報パケットサイズと関係する受信ホスト要求に関係なく情報パケットを転送する、請求項１に記載の装置。

【請求項４】前記リセグメンテーションエンティティが、中間の転送エンティティである、請求項１に記載の装置。

【請求項５】送信ホストがただ、サブセグメントを包含するＩＰデータグラムが損失する事象において、損失され、完全なセグメントでないＴＣＰデータを転送しなければならない、請求項１に記載の装置。

【請求項６】前記ホストの各々がＩＰホストであり、前記ホストの各々が、転送プロトコルとしてＴＣＰを使用する、請求項１に記載の装置。

【請求項７】前記ホストのうちの少なくとも１つが、１又はそれ以上のアプリケーションと、オペレーティングシステムと、ＴＣＰプロトコルエンティティと、ＩＰプロトコルエンティティと、１又はそれ以上のネットワークデバイスドライバとを含むシステムを有する、請求項１に記載の装置。

【請求項８】前記各ホストのＭＴＵが、転送及び受信目的の両方に関して関係が無く、それ故、多くの異種のシステム及びネットワークを有する異種事業形態が、システム性能を低下させることなく、若しくは、エンタープライズ広域通信を許容する特別な調整を備えるホストを妨げることなく提供される、請求項１に記載の装置。

【請求項９】前記リセグメントエンティティが、転送エンジンと、ＴＣＰリセグメンターと、ＩＰ層と、１又はそれ以上のネットワークインターフェースとを有する、請求項１に記載の装置。

【請求項１０】前記リセグメントエンティティが、ルータ、ブリッジ、スイッチ、又は、送信ホストのドライバ若しくはネットワークインターフェースのうちのいずれかである、請求項１に記載の装置。

【請求項１１】前記オリジナル情報パケットが、ＩＰヘッダと、ＴＣＰヘッダと、ＴＣＰデータとを含むＩＰデータグラムである、請求項１に記載の装置。

【請求項１２】各リセグメントされたパケットが、ＩＰヘッダと、リセグメントされたＴＣＰヘッダと、リセグメントされたＴＣＰデータとを含むＩＰデータグラムである、請求項１に記載の装置。

【請求項１３】前記リセグメントエンティティが、ＴＣＰではない転送プロトコルを包含する情報パケットでＩＰフラグメンテーションを実行する、請求項１に記載の装置。

【請求項１４】電子ネットワークにわたる２又はそれ以上のホストの間で情報パケットを交換するためのリセグメンテーションエンティティであって、少なくとも前記ホストのうちの１つが、前記２又はそれ以上のホストの他のものとサイズが異なる情報パケットを転送し且つ受信し、前記情報パケット内の大きな情報セグメントを、複数の対応する情報サブパケット内の複数の小さなサブセグメントにリセグメントするために、前記ホストの前記１又はそれ以上のものか、ネットワークインターフェースのいずれかの内部に、又は、前記電子ネットワークの内部であり、前記ホストから離れて配置されたリセグメンターを有する、リセグメンテーションエンティティ。

【請求項１５】電子ネットワークにわたって２つ又はそれ以上のホストの間で情報パケットを交換するための方法であって、前記ホストのうちの少なくとも１つが、前記２又はそれ以上のホストの他のものとサイズが異なる情報パケットを転送し且つ受信し、前記情報パケット内の大きな情報セグメントを、複数の対応する情報サブパケット内の複数の小さなサブセグメントにリセグメントし前記情報サブパケットを受信ホストに転送する、ステップを有する方法。

【請求項１６】受信ホストは、情報パケットがまるでそれらが前記受信ホストに関して特に転送されるように見えるサイズのものである前記情報パケットを受信する、請求項１５に記載の方法。

【請求項１７】送信ホストが、情報パケットサイズと関係する受信ホスト要求に関係なく情報パケットを転送する、請求項１５に記載の方法。

【請求項１８】前記リセグメント・ステップが、中間の転送エンティティを含むリセグメンテーションエンティティによって実行される、請求項１５に記載の方法。

【請求項１９】送信ホストがただ、リセグメントされたセグメントを包含するパケットが損失する事象において、損失され、完全なセグメントでないＴＣＰデータを転送しなければならない、請求項１５に記載の方法。

【請求項２０】前記ホストの各々がＩＰホストであり、前記ホストの各々が、転送プロトコルとしてＴＣＰを使用する、請求項１５に記載の方法。

【請求項２１】前記ホストのうちの少なくとも１つが、

１又はそれ以上のアプリケーションと、オペレーティングシステムと、ＴＣＰプロトコルエンティティと、ＩＰプロトコルエンティティと、１又はそれ以上のネットワークデバイスドライバとを含むシステムを有する、請求項１５に記載の方法。

【請求項２２】前記各ホストのＭＴＵが、転送及び受信目的の両方に関して関係が無く、それ故、多くの異種のシステム及びネットワークを有する異種事業形態が、システム性能を低下させることなく、若しくは、エンタープライズ広域通信を許容する特別な調整を備えるホストを妨げることなく提供される、請求項１５に記載の方法。

【請求項２３】前記リセグメントエンティティが、転送エンジンと、ＴＣＰリセグメンターと、ＩＰプロトコルエンティティと、１又はそれ以上のネットワークインターフェースとを有する、請求項１８に記載の方法。

【請求項２４】前記リセグメントエンティティが、ルータ、ブリッジ、スイッチ、又は、送信ホストのドライバ若しくはネットワークインターフェースのうちのいずれかである、請求項１８に記載の方法。

【請求項２５】前記情報パケットが、ＩＰヘッダと、ＴＣＰヘッダと、ＴＣＰデータとを含むＩＰデータグラムである、請求項１５に記載の方法。

【請求項２６】各サブパケットが、ＩＰヘッダと、リセグメントされたＴＣＰヘッダと、リセグメントされたＴＣＰデータとを含むＩＰデータグラムである、請求項１５に記載の方法。

【請求項２７】前記リセグメントエンティティが、ＴＣＰではない転送プロトコルを包含する情報パケットでＩＰフラグメンテーションを実行する、請求項１８に記載の方法。

【請求項２８】電子ネットワークにわたる２又はそれ以上のホストの間で情報パケットを交換するためのリセグメンテーション方法であって、少なくとも前記ホストのうちの１つが、前記２又はそれ以上のホストの他のものとサイズが異なる情報パケットを転送し且つ受信し、
前記電子ネットワークの内部であり、前記ホストから離れているか、若しくは、前記ホスト又はネットワークインターフェースの内部のいずれかに、配置されたリセグメンターを提供し、
前記情報パケット内の大きな情報セグメントを、複数の対応する情報サブパケット内の複数の小さなサブセグメントにリセグメントする、ステップを有する前記リセグメンテーション方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は電子ネットワークに関する。特に、本発明は、異なるＭＴＵｓを有し、それらのトランスポートプロトコルとしてＴＣＰを主に使用する、２つのＩＰの間の有効な通信を確立し、維持することに関する。

【０００２】

【従来の技術】ＩＰ（インターネット・プロトコル）ルーティング・エンティティが、異なるＭＴＵｓ（最大トランスミッション・ユニット）を備えるインターフェースを有し、ＴＣＰ（トランスミッション制御プロトコル）セグメントを包含する大きなＩＰパケットが、小さなＭＴＵを有するインターフェースから転送されなければならないとき、ルーティング・エンティティは典型的には、以下の２つのうちの１つのことをなす：

・ＩＰフラグメンテーションを使用してパケットを分解する；又は

・パケットをドロップし、ＩＣＭＰ（インターネット制御メッセージプロトコル）ＤＥＳＴＩＮＡＴＩＯＮ ＵＮＲＥＡＣＨＡＢＬＥを送り；ＦＲＡＧＭＥＮＴＡＴＩＯＮ ＮＥＥＤＥＤメッセージを送信ホストにバックする。

【０００３】ＩＰフラグメンテーションは、大きなＩＰパケットを、受信ホストによってリアセンブリを要求する種々の小さなＩＰフラグメントに分解する。パケットをドロップすることによりデータは失われ、ＩＣＭＰ ＤＥＳＴＩＮＡＴＩＯＮ ＵＮＲＥＡＣＨＡＢＬＥを送信し；ＦＲＡＧＭＥＮＴＡＴＩＯＮ ＮＥＥＤＥＤメッセージをホストにバックさせることにより、ホストはそのパスＭＴＵを減少させ、より小さなＩＰパケットを使用することができる。

【０００４】ＩＰフラグメンテーションは、ＩＰフラグメントをリアセンブリする負担を受信ホストに置く。ホストがＩＰフラグメントを受信するとき、それがペイロードをＴＣＰに配信する前に、それが完全なＩＰパケットを形成するように全てのフラグメントを完全にリアセンブリしなければならない。このことは、追加のバッファとＣＰＵリソースを受信ホストで要求し、ＴＣＰセグメントを受信する待ち時間を増加させる。更に、どんなフラグメントがドロップし、さもなければ、通過中ならば、全体のオリジナルＴＣＰセグメントは再転送されなければならない。

【０００５】ＩＰパケットをドロップし、ＩＣＭＰ ＤＥＳＴＩＮＡＴＩＯＮ ＵＮＲＥＡＣＨＡＢＬＥを送信し-ＦＲＡＧＭＥＮＴＡＴＩＯＮ ＮＥＥＤＥＤメッセージをホストにバックさせることにより、ホストのパスＭＴＵ発見アルゴリズムはパスに関する小さなＭＴＵを選択することができ、将来の転送に関しより小さなＩＰパケットを使用することができる。これが生じたとき、それは、ホストの効率を低下させ、より小さなＭＴＵを使用するためにより大きなＭＴＵを使用する能力があるホストを強制する。それはまた、再転送されるべき全体のＴＣＰセグメントを要求する。

【０００６】大きなＩＰパケットを運ぶことができないパスにわたって通信するときでさえ、大きなＩＰパケッ

トを送信することができるホストがそのようにすることができる利点がある。それはまた、種々のＩＰフラグメントからのＩＰパケットをリアセンブリするタスクを備える受信ホストに負担をかけないという利点がある。最後に、パケットがそれがネットワークを横切るようにとるパスに沿って、インターフェースの最小のＭＴＵを使用するようにネットワークにおいて、全ての結合されたステーションを要求することなく、この能力を可能にするという利点がある。

【０００７】

【課題を解決するための手段】本発明は、ＴＣＰリセグメンテーション技術を実施するリセグメンテーションエンティティを提供し、受信ホストは、それがまるで受信ホストのＭＴＵに関して特に転送されるように見えるパケットを受信する。受信ホストは、ＩＰリアセンブリのために必要なＣＰＵ利用及びバッファリングを要求しない。また、受信ホストは、ＴＣＰセグメントを処理する前に、それがフラグメントからのＩＰデータグラムをリアセンブリするのに必要ならば、それがなされるよりも、リセグメントされたＴＣＰセグメントを包含するＩＰデータグラムを受信するとき、短い待ち時間を有する。更に、送信ホストは、受信ステーションのＭＴＵに関係なく、その最大ＭＴＵでＴＣＰセグメントを転送し、中間のルーティングエンティティはＴＣＰリセグメンテーションが生じることを保証することが知られている。リセグメントされたＴＣＰセグメントを包含するＩＰデータグラムが失われる事象の際、送信ホストは、損失し、完全なＴＣＰセグメントでない、実際のＴＣＰデータを転送しなければならないだけである。

【０００８】

【発明の実施の形態】本発明は、新規なＴＣＰリセグメンテーション技術を実行するリセグメンテーションエンティティを提供し、受信ホストが、まるでそれらが受信ホストのＭＴＵに関して特に転送されているかのように見えるパケットを受信する。受信ホストは、ＩＰリアセンブリのために必要なＣＰＵ利用及びバッファを要求しない。また、受信ホストは、ＴＣＰセグメントを処理する前に、それがフラグメントからのＩＰデータグラムをリアセンブリするのに必要ならば、それがなされるよりも、リセグメントされたＴＣＰセグメントを包含するＩＰデータグラムを受信するとき、短い待ち時間を有する。更に、送信ホストは、受信ステーションのＭＴＵに関係なく、そのローカルＭＴＵでＴＣＰセグメントを転送する。リセグメントされたＴＣＰセグメントを包含するＩＰデータグラムが失われる事象の際、送信ホストは、損失し、完全なＴＣＰセグメントでない、実際のＴＣＰデータを転送しなければならないだけである。

【０００９】図１は、ホストプロトコルスタックを提供するシステムを示すブロック概念図である。図１に示したように、システム１０は、１又はそれ以上のアプリケーション１２と、オペレーティングシステム１３と、ＴＣＰプロトコルエンティティ１４と、ＵＤＰ（User Datagram Protocol：ユーザデータグラムプロトコル）プロトコルエンティティ１５と、ＩＰプロトコルエンティティ１６と、１又はそれ以上のネットワークデバイスドライバ１７，１８，１９とを含む。本発明は、図１に示したような、あらゆるシステムに容易に適用される。ホストは、受信ホスト又は転送ホストのいずれであってもよい。本発明の一意的な態様は、ホストのＭＴＵサイズが、転送及び受信目的の両方に関係がないことである。それ故、多くの本質的に異なるシステム及びネットワークを有する異種事業形態が、システムパフォーマンスの低下、又は、エンタープライズ広域通信を許容するための特別な適応を備えるホストを妨げることなく提供されうる。

【００１０】図２は、本発明による、ルータ２１の形態での、リセグメントエンティティを示すブロック概略図である。リセグメントエンティティはまた、転送エンジン２０と、ＴＣＰリセグメンター２２と、ＩＰプロトコルエンティティ１６と、１又はそれ以上のネットワークインターフェース２７，２８，２９とを含む。リセグメンテーションエンティティは、ここでは例示の目的だけのためにルータ内に示す。かかるエンティティは、いかなるネットワーク要素、例えば、ルータ、ブリッジ、スイッチ、又は、送信ホストのドライバ若しくはネットワークインターフェースを含みうる。

【００１１】図３は、本発明によるホスト通信を示すブロック概略図である。図３では、第１のシステム３０が小さなＭＴＵネットワーク３４にあり、第２のシステム３２が大きなＭＴＵネットワーク３５にある。２つのネットワークに配置されたシステムは、図２のリセグメントエンティティを含むルータ２１を介して情報を交換することができる。

【００１２】図４は、ＩＰヘッダ４５と、ＴＣＰヘッダ４６と、ＴＣＰデータ４７とを包含するＴＣＰ／ＩＰパケット４４の概略を示す。ＩＰフラグメンテーション、又は、従来技術のようなＩＰフラグメンテーションの無効をあてにするだけの代わりに、本発明は、ＩＰフラグメンテーションとＴＣＰリセグメンテーションとの組み合わせを使用するシステムを提供する。これにより、ＩＰ及びＴＣＰプロトコルエンティティは、全ての接続されたネットワーク及びホストが大きなＭＴＵをサポートする要求なしで、より大きなＭＴＵを使用した結果から生じる効率を増加させる利益を得ることができる。

【００１３】従って、本発明は、単一のＴＣＰセグメントを得、それをサブセグメントと呼ばれる多数のＴＣＰセグメントに分解する、ＴＣＰリセグメンテーションと呼ばれる機構を提供する。このリセグメンテーションは、ネットワーク要素（ルータ、ブリッジ、スイッチ）又は、送信ホストのドライバ又はネットワークインター

フェースによって中間でホップを生じる。

【００１４】図５は、本発明によるリセグメントされたパケットの概略を示す。従って、（図４に示した）オリジナルＩＰデータグラム４４におけるオリジナルセグメント４７は、多数のサブセグメント５２，５３，５４に分解され、その各々は、完全なＩＰデータグラム６１，６３，６５に含まれており、各々ＩＰヘッダ４５ａ，４５ｂ，４５ｃ及びリセグメントされたＴＣＰヘッダ５１ａ，５１ｂ，５１ｃを包含する。リセグメンテーションは、ＴＣＰリセグメンテーションと関係して実際には実行される。

【００１５】図６は、本発明による電子ネットワークにおけるリセグメンテーションを示すブロック概略図である。本発明の好ましい実施形態では、ＴＣＰリセグメンテーションは以下のように作動する：ＩＰルーティング、ブリッジ、スイッチングエンティティ２１が異なるＭＴＵを備えるネットワークインターフェース２７，２８，２９を有し、転送ステーション６２からの大きなＩＰデータグラム４４が小さなＭＴＵを有するインターフェース２７の外に転送されなければならないとき、ＩＰデータグラムは完全なＴＣＰセグメントを包含し、パケットは、ＴＣＰリセグメントアルゴリズムに関係する複数の小さなＴＣＰサブセグメント６１，６３，６５にリセグメントされる。このことにより、受信ステーション６０は、ＩＰリアセンブリに関する受信ステーションのＣＰＵリソース又は追加のバッファを消費することなしに、ＴＣＰサブセグメントの各々を独立に処理することができる。

【００１６】ＴＣＰサブセグメントが失われるならば、ＴＣＰストリームの全てのバイトが、ナンバリングされ、独立して応答されるので、送信ホスト６２は、失われ、完全なＴＣＰセグメントではないＴＣＰデータを再転送するだけのために必要である。

【００１７】ＴＣＰリセグメンテーションは、ＩＰ ＤＯＮ’Ｔ ＦＲＡＧＭＥＮＴビットの設定を無視し、変更しうる。従って、ＭＴＵ発見アルゴリズムは、仕事がローカルインターフェースのＭＴＵである最大ＭＴＵ、及び、パスの毎リンクによって使用されうる必然的でない最小ＭＴＵを報告することによって失敗しうる。

【００１８】ＴＣＰリセグメンテーションは、リセグメントエンティティが、ＴＣＰではない転送プロトコルを包含するパケット、及び、リセグメンテーションに関する基準を満たさないＴＣＰパケットのＩＰフラグメンテーションを実行することを要求する。

【００１９】各新しいセグメントで新しいＴＣＰ検査合計の生成を要求するので、ＴＣＰリセグメンテーションは、プロセッサ集中オペレーションである。このため、ＴＣＰリセグメンテーションが、可能なときはいつでもハードウェア支援で実行されることが好ましいが、要求されない。

【００２０】セグメントに分けられたパケットの発信者は、断片で受信されるように、セグメントの部分的なＡＣＫを受信しうる。これはＴＣＰプロトコル定義において許されるけれども、幾つかのＴＣＰプロトコル実行は、これを正確に取り扱えない。

【００２１】以下は、フラグメントされておらず、いかなるＩＰ又はＴＣＰオプションをも包含しないＴＣＰセグメントをリセグメントするアルゴリズムの疑似コード例である。それは、ＩＰ ＴＴＬフィールドを減少させず、ＩＰ ＤＯＮ’Ｔ ＦＲＡＧＭＥＮＴビットを無視する。それはまた、ＴＣＰリセグメンテーションを実行することができないパケットに関して、ＩＰ ＤＯＮ’Ｔ ＦＲＡＧＭＥＮＴビットをターンオフさせ、全てのリセグメントされたパケットに関してＩＰ ＤＯＮ’Ｔ ＦＲＡＧＭＥＮＴビットをターンオンする。

【００２２】

アルゴリズム

```
IF the received packet is IP AND
        the IP length exceeds the MTU AND
        no IP options are present AND
        the IP datagram has not already been fragmented AND
        the protocol is TCP AND
        no TCP options are present THEN
    set the remaining data length to the original data length
    set the current data pointer to the original data pointer
    set the new urgent pointer to the original urgent pointer
    set the current sequence number to the original sequence
        number
    WHILE the remaining data length is non-zero DO
        allocate a new transmit buffer for the new segment
        set the new segment length to the minimum of the length
```

```
                of the remaining data and the egress MSS
            copy the original IP and TCP headers to the new transmit
                buffer
            IF the TCP SYN bit is set AND
                        this is not the first subsegment THEN
                turn off the TCP SYN bit
            IF the TCP FIN bit is set AND
                        this is not the last subsegment THEN
                turn off the TCP FIN bit
            IF the TCP PUSH bit is set AND
                        this is not the last subsegment THEN
                turn off the TCP PUSH bit
            IF the TCP URG bit is set AND
                        the new urgent pointer is non-zero THEN
                set the urgent pointer to the new urgent pointer
                IF the new urgent pointer points to within the
                            subsegment THEN
                        set the new urgent pointer to zero
                ELSE
                        decrement the new urgent pointer by the
                            subsegment length
            ELSE
                turn off the TCP URG bit
                set the urgent pointer to zero
            set the TCP sequence number to the current sequence
                number
            increment the current sequence number by the
                subsegment length
            set the IP datagram length to the new IP datagram length
            turn on the IP DONT FRAGMENT bit
            decrement the remaining data length by the
                subsegment length
            copy the data from the current data pointer to
                the data portion of the new transmit buffer for the
                length of the subsegment
            increment the current data pointer by the
                subsegment length
            recalculate the TCP checksum
            recalculate the IP checksum
            send the new segment
        ENDWHILE
        free the original data buffer
    ELSE
        turn off the IP DONT FRAGMENT bit
```

【００２３】本発明を好ましい実施形態を参照してここに記載したけれども、当業者にとって、本発明の精神及び範囲から逸脱することなくここで説明したものを他のアプリケーションで置換することが容易であることは明らかである。従って、本発明は特許請求の範囲によってのみ限定されるべきである。

【図面の簡単な説明】

【図１】ホストプロトコルスタックを示すブロック概略図である。

【図２】本発明によるリセグメントエンティティを示すブロック概略図である。

【図３】本発明によるホスト通信を示すブロック概略図である。

【図４】ＴＣＰ／ＩＰパケットの概略図である。

【図５】本発明によるリセグメントされたパケットの概略図である。

【図６】本発明による電子ネットワークにおけるリセグメンテーションを示すブロック概略図である。

【符号の説明】

１２　アプリケーション

１３　オペレーティングシステム

１４　ＴＣＰ

１５　ＵＤＰ

１６　ＩＰ

２０　ルータ

２２　リセグメンター

２７、２８、２９　インターフェース

３４　小ＭＴＵネットワーク

３５　大ＭＴＵネットワーク

４５　ＩＰヘッダ

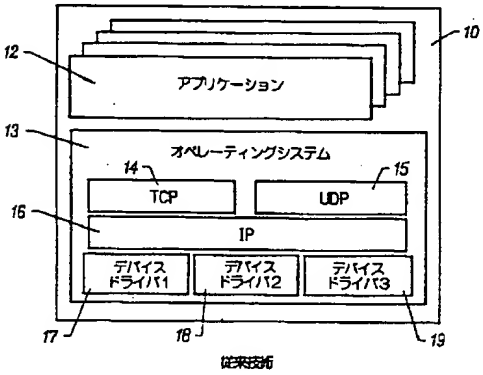
４６　ＴＣＰヘッダ

４７　セグメント

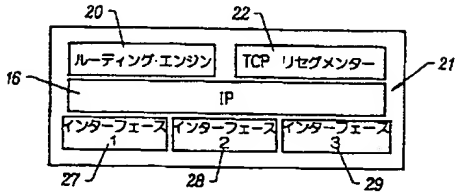
５１　リセグメントされたＴＣＰヘッダ

５２、５３、５４　サブセグメント

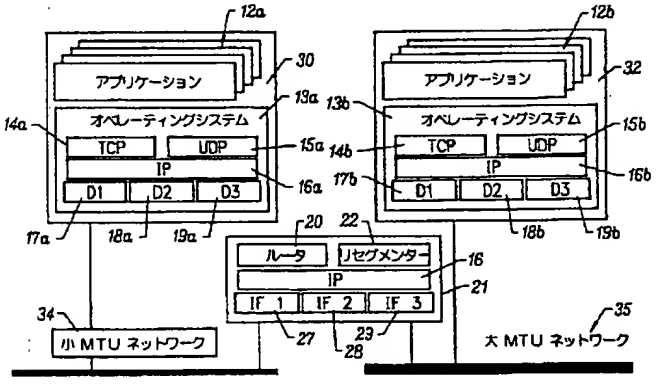
【図１】



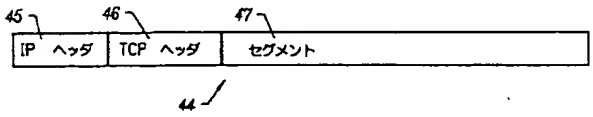
【図２】



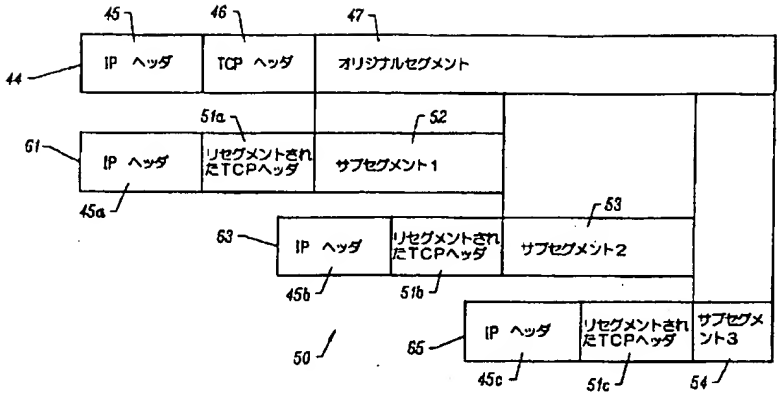
【図３】



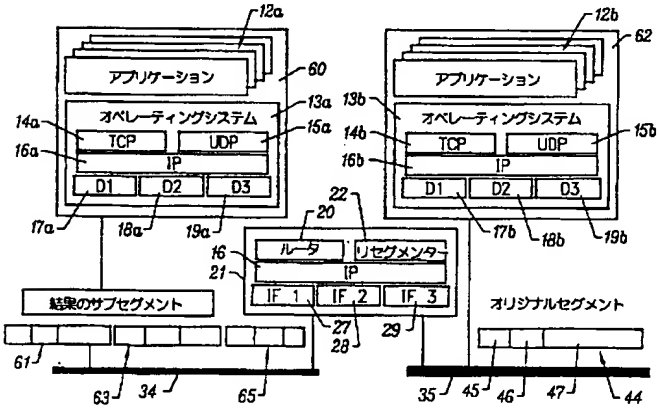
【図４】



【図５】



【図６】



フロントページの続き

(72)発明者　ジョン　ヘイズ
アメリカ合衆国　カリフォルニア州
95065　サンタ　クルーズ　ジャーヴィス
ロード　315

(72)発明者　ウェイン　ハザウェイ
アメリカ合衆国　カリフォルニア州
94087　サニーヴェイル　カシミア　コート　535

【外国語明細書】

TCP RESEGMENTATION

BACKGROUND OF THE INVENTION

TECHNICAL FIELD

The invention relates to electronic networks. More particularly, the invention relates to establishing and maintaining efficient communication between two IP hosts that have differing MTUs and that primarily use TCP as their transport protocol.

DESCRIPTION OF THE PRIOR ART

When an IP (Internet Protocol) routing entity has interfaces with differing MTUs (Maximum Transmission Unit) and a large IP packet containing a TCP (Transmission Control Protocol) segment must be forwarded from an interface having a smaller MTU, the routing entity typically does one of two things:

- Fragment the packet using IP fragmentation; or
- Drop the packet and send an ICMP (Internet Control Message Protocol) DESTINATION UNREACHABLE: FRAGMENTATION NEEDED message back to the sending host.

IP fragmentation breaks the large IP packet into several smaller IP fragments which require reassembly by the receiving host. Dropping the packet causes the data to be lost, and sending an ICMP DESTINATION UNREACHABLE;

FRAGMENTATION NEEDED message back to the host causes the host to reduce its path MTU and use smaller IP packets.

IP fragmentation places the burden of reassembling the IP fragments onto the receiving host. When a host receives IP fragments, it must fully reassemble all of the fragments to form a complete IP packet before it can deliver the payload to TCP. This requires additional buffer and CPU resources at the receiving host and increases the latency of receiving a TCP segment. Additionally, if any of the fragments are dropped or otherwise lost during transit, the entire original TCP segment must be retransmitted.

Dropping the IP packet and sending an ICMP DESTINATION UNREACHABLE - FRAGMENTATION NEEDED message back to the sending host causes the host's path MTU discovery algorithm to choose a smaller MTU for the path and use that smaller MTU for future transmissions. When this occurs, it forces a host that is capable of using a larger MTU to use a smaller MTU, reducing the host's efficiency. It also requires the entire TCP segment to be retransmitted.

It would be advantageous to allow a host that is capable of sending large IP packets to do so, even when it is communicating over a path that is not capable of carrying large IP packets. It would also be advantageous not to burden the receiving host with the task of reassembling the IP packet from the various IP fragments. Finally, it would be advantageous to enable this ability without requiring all of the connected stations in the network to use the smallest MTU of any interface along the path that a packet takes as it traverses the network.

SUMMARY OF THE INVENTION

The invention provides a resegmentation entity that implements a TCP resegmentation technique wherein a receiving host receives packets that appear as if it they have been transmitted specifically for the receiving host's MTU. The receiving host does not require the buffering and CPU utilization necessary for IP reassembly. Also, the receiving host has a lower latency when receiving IP datagrams that contain resegmented TCP segments than it would if it needed to re-assemble an IP datagram from fragments before it could process the TCP segment. Further, the sending host transmits TCP segments at its largest MTU, without regard to the receiving station's MTU, knowing that the intermediate routing entity insures that TCP resegmentation occurs. In the event that an IP datagram containing a re-segmented TCP segment is lost, the sending host only has to retransmit the actual TCP data that was lost, and not the complete TCP segment.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block schematic diagram showing a host protocol stack;

Fig. 2 is a block schematic diagram showing a resegmenting entity according to the invention;

Fig. 3 is a block schematic diagram showing host communications according to the invention;

Fig. 4 is a schematic representation of a TCP/IP packet;

Fig. 5 is a schematic representation of a resegmented packet according to the invention; and

Fig. 6 is a block schematic diagram showing resegmentation in an electronic network according to the invention.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides a resegmentation entity that implements a novel TCP resegmentation technique wherein a receiving host receives packets that appear as if it they have been transmitted specifically for the receiving host's MTU. The receiving host does not require the buffering and CPU utilization necessary for IP reassembly. Also, the receiving host has a lower latency when receiving IP datagram that contain resegmented TCP segments than it would if it needed to reassemble an IP datagram from fragments before it could process the TCP segment. Further, the sending host transmits TCP segments at its local MTU, without regard to the receiving station's MTU. In the event that an IP datagram containing a resegmented TCP segment is lost, the sending host only has to retransmit the actual TCP data that was lost, and not the complete TCP segment.

Fig. 1 is a block schematic diagram showing a system that provides a host protocol stack. As shown in Fig. 1, a system 10 includes one or more applications 12, an operating system 13, a TCP protocol entity 14, a UDP (User Datagram Protocol) protocol entity 15, an IP protocol entity 16, and one or more network device drivers 17, 18, 19. The invention is readily applied to any system, such as that shown in Fig. 1. The host may be either a receiving host or a transmitting host. A unique aspect of the invention is that the MTU size of the host is irrelevant for both transmitting and receiving purposes. Therefore, a heterogeneous enterprise having many disparate systems and networks may

be provided without degrading system performance or encumbering a host with special adaptations to allow enterprise wide communication.

Fig. 2 is a block schematic diagram showing a resegmenting entity, in the form of a router 21, according to the invention. The resegmenting entity also includes a forwarding engine 20, a TCP resegmenter 22, an IP protocol entity 16, and one or more network interfaces 27, 28, 29. The resegmenting entity is shown herein within a router for purposes of example only. Such entity may be included in any network elements, *e.g.* a router, bridge, switch, or a sending host's driver or network interface.

Fig. 3 is a block schematic diagram showing host communications according to the invention. In Fig. 3, a first system 30 resides on a small MTU network 34 and a second system 32 resides on a large MTU network 35. Systems located on the two networks are able to exchange information via a router 21 which comprises the resegmenting entity of Fig. 2.

Fig. 4 is a schematic representation of an TCP/IP packet 44 which includes an IP header 45, a TCP header 46, and TCP data 47. Instead of relying only on IP fragmentation, or the avoidance of IP fragmentation as in the prior art, the invention provides a system that uses a combination of IP fragmentation and TCP resegmentation. This allows IP and TCP protocol entities to take advantage of the increase in efficiency that results from using larger MTUs, without requiring that all connected networks and hosts support large MTUs.

Thus, the invention provides a mechanism referred to as TCP resegmentation, which takes a single TCP segment and breaks it into multiple TCP segments, called subsegments. This resegmentation occurs at an intermediate hop by a

network element (router, bridge, switch) or in the sending host's driver or network interface.

Fig. 5 is a schematic representation of a resegmented packet according to the invention. Thus, the original segment 47 in the original IP datagram 44 (shown in Fig. 4) is broken up in multiple subsegments 52, 53, 54, each of which is contained in a complete IP datagram 61, 63, 65, each including an IP header 45a, 45b, 45c and a resegmented TCP header 51a, 51b, 51c. The resegmentation is actually performed in accordance with the TCP resegmentation.

Fig. 6 is a block schematic diagram showing resegmentation in an electronic network according to the invention. In the preferred embodiment of the invention, TCP resegmentation operates as follows:

When an IP routing, bridging, switching entity 21 has network interfaces 27, 28, 29 with differing MTUs and a large IP datagram 44 from a transmitting station 62 must be forwarded out an interface 27 having a smaller MTU, where the IP datagram contains a complete TCP segment, the packet is resegmented into multiple, smaller TCP subsegments 61, 63, 65 in accordance with the TCP resegmenting algorithm. This allows the receiving station 60 to process each of the TCP subsegments independently, without consuming additional buffer or CPU resources on the receiving station for IP reassembly.

If a TCP subsegment is lost, the sending host 62 needs to only retransmit the TCP data that was lost, not the complete TCP segment, because all bytes in a TCP stream are numbered and acknowledged individually.

TCP resegmentation ignores and may change the setting of the IP DON'T FRAGMENT bit. Accordingly, MTU discovery algorithms may fail by reporting

that the largest MTU that works is the MTU of the local interface, and not necessarily the smallest MTU that can be used by every link in the path.

TCP resegmentation requires that the resegmenting entity perform IP fragmentation on packets which contain transport protocols other than TCP and on TCP packets that do not meet the criteria for resegmentation.

TCP resegmentation is a processor intensive operation because it requires the generation of a new TCP checksum on each new segment. Because of this, it is preferred, but not required that TCP resegmentation be performed with hardware assistance whenever possible.

The originator of the segmented packet may receive partial ACKs of the segment as it is received in pieces. Although this is permissible in the TCP protocol definition, some TCP protocol implementations may not deal with this correctly.

The following is a pseudocode example of an algorithm that resegments TCP segments that have not been fragmented and that do not contain any IP or TCP options. It does not decrement the IP TTL field and it ignores the IP DON'T FRAGMENT bit. It also turns off the IP DON'T FRAGMENT bit for any packets on which it is unable to perform TCP resegmentation, and it turns on the IP DON'T FRAGMENT bit for all resegmented packets.

Algorithm

```
IF the received packet is IP AND
        the IP length exceeds the MTU AND
        no IP options are present AND
        the IP datagram has not already been fragmented AND
        the protocol is TCP AND
        no TCP options are present THEN

    set the remaining data length to the original data length
    set the current data pointer to the original data pointer
    set the new urgent pointer to the original urgent pointer
    set the current sequence number to the original sequence
        number

    WHILE the remaining data length is non-zero DO
        allocate a new transmit buffer for the new segment
        set the new segment length to the minimum of the length
            of the remaining data and the egress MSS
        copy the original IP and TCP headers to the new transmit
            buffer

        IF the TCP SYN bit is set AND
                this is not the first subsegment THEN
            turn off the TCP SYN bit
        IF the TCP FIN bit is set AND
                this is not the last subsegment THEN
            turn off the TCP FIN bit
```

```
IF the TCP PUSH bit is set AND
        this is not the last subsegment THEN
    turn off the TCP PUSH bit
IF the TCP URG bit is set AND
        the new urgent pointer is non-zero THEN
    set the urgent pointer to the new urgent pointer
    IF the new urgent pointer points to within the
            subsegment THEN
        set the new urgent pointer to zero
    ELSE
        decrement the new urgent pointer by the
            subsegment length
ELSE
    turn off the TCP URG bit
    set the urgent pointer to zero

set the TCP sequence number to the current sequence
    number
increment the current sequence number by the
    subsegment length

set the IP datagram length to the new IP datagram length
turn on the IP DONT FRAGMENT bit
decrement the remaining data length by the
    subsegment length
copy the data from the current data pointer to
    the data portion of the new transmit buffer for the
    length of the subsegment
```

```
            increment the current data pointer by the
                subsegment length
            recalculate the TCP checksum
            recalculate the IP checksum
            send the new segment
        ENDWHILE
        free the original data buffer
ELSE
        turn off the IP DONT FRAGMENT bit
```

Although the invention is described herein with reference to the preferred embodiment, one skilled in the art will readily appreciate that other applications may be substituted for those set forth herein without departing from the spirit and scope of the present invention. Accordingly, the invention should only be limited by the Claims included below.

CLAIMS

1. An apparatus for exchanging information packets between two or more hosts over an electronic network, wherein at least one of said hosts transmits and receives information packets that differ in size from those of the other of said two or more hosts, comprising:

a resegmentation entity located within either of said one or more of said hosts or their respective network interfaces or within said electronic network, for resegmenting larger information segments within said information packets into a plurality of smaller subsegments within a plurality of corresponding information subpackets.

2. The apparatus of Claim 1, wherein a receiving host receives information packets that are of a size wherein said information packets appear as if it they have been transmitted specifically for said receiving host.

3. The apparatus of Claim 1, wherein a sending host transmits information packets without regard to a receiving hosts requirements with regard to information packet size.

4. The apparatus of Claim 1, wherein said resegmentation entity is an intermediate forwarding entity.

5. The apparatus of Claim 1, wherein a sending host only has to retransmit the TCP data that was lost, and not a complete segment in the event that an IP datagram containing a subsegment is lost.

6. The apparatus of Claim 1, wherein each of said hosts is an IP host, and wherein each of said hosts use TCP as a transport protocol.

7. The apparatus of Claim 1, wherein at least one of said hosts comprises a system that includes one or more applications, an operating system, a TCP protocol entity, an IP protocol entity, and one or more network device drivers.

8. The apparatus of Claim 1, wherein the MTU of each said host is irrelevant for both transmitting and receiving purposes and, therefore, a heterogeneous enterprise having many disparate systems and networks may be provided without degrading system performance or encumbering a host with special adaptations to allow enterprise wide communication.

9. The apparatus of Claim 1, wherein said resegmenting entity comprises a forwarding engine, a TCP resegmenter, an IP layer, and one or more network interfaces.

10. The apparatus of Claim 1, wherein said resegmenting entity is any of a router, bridge, switch, or a sending host's driver or network interface.

11. The apparatus of Claim 1, wherein said original information packet is an IP datagram which includes an IP header, a TCP header, and TCP data.

12. The apparatus of Claim 1, wherein each resegmented packet is an IP datagram which includes an IP header, a resegmented TCP header, and resegmented TCP data.

13. The apparatus of Claim 1, wherein said resegmenting entity performs IP fragmentation on information packets which contain transport protocols other than TCP.

14. A resegmentation entity for exchanging information packets between two or more hosts over an electronic network, wherein at least one of said hosts transmits and receives information packets that differ in size from those of the other of said two or more hosts, said resegmentation entity comprising:

a resegmentor, located within either of one or more of said hosts or network interfaces or within said electronic network and apart from said hosts, for resegmenting larger information segments within said information packets into a plurality of smaller subsegments within a plurality of smaller subsegments within a plurality of corresponding information subpackets .

15. A method for exchanging information packets between two or more hosts over an electronic network, wherein at least one of said hosts transmits and receives information packets that differ in size from those of the other of said two or more hosts, said method comprising the steps of:

resegmenting larger information segments within said information packets into a plurality of smaller subsegments within a plurality of corresponding information subpackets; and

forwarding said information subpackets to a receiving host.

16. The method of Claim 15, wherein said receiving host receives information packets that are of a size wherein said information packets appear as if it they have been transmitted specifically for said receiving host.

17. The method of Claim 15, wherein a sending host transmits information packets without regard to a receiving hosts requirements with regard to information packet size.

18. The method of Claim 15, wherein said resegmenting step is implemented by a resegmentation entity comprising an intermediate forwarding entity.

19. The method of Claim 15, wherein a sending host only has to retransmit the TCP data that was lost, and not a complete segment in the event that a packet containing a resegmented segment is lost.

20. The method of Claim 15, wherein each of said hosts is an IP host, and wherein each of said hosts use TCP as a transport protocol.

21. The method of Claim 15, wherein at least one of said hosts comprises a system that includes one or more applications, an operating system, a TCP protocol entity, an IP protocol entity, and one or more network device drivers.

22. The method of Claim 15, wherein the MTU of each said host is irrelevant for both transmitting and receiving purposes and, therefore, a heterogeneous enterprise having many disparate systems and networks may be provided without degrading system performance or encumbering a host with special adaptations to allow enterprise wide communication.

23. The method of Claim 18, wherein said resegmenting entity comprises a forwarding engine, a TCP resegmenter, an IP protocol entity, and one or more network interfaces.

24. The method of Claim 18, wherein said resegmenting entity is any of a router, bridge or switch, or a sending host's driver or network interface.

25. The method of Claim 15, wherein said information packet is an IP datagram which includes an IP header, a TCP header, and TCP data.

26. The method of Claim 15, wherein each subpacket is a IP datagram which includes an IP header, a resegmented TCP header, and resegmented TCP data.

27. The method of Claim 18, wherein said resegmenting entity performs IP fragmentation on information packets which contain transport protocols other than TCP.

28. A resegmentation method for exchanging information packets between two or more hosts over an electronic network, wherein at least one of said hosts transmits and receives information packets that differ in size from those of the other of said two or more hosts, said resegmentation method comprising the steps of:

providing a resegmenter located either of within said electronic network and apart from said hosts or within said hosts or network interfaces; and

resegmenting larger information segments within said information packets into a plurality of smaller subsegments within a plurality of corresponding information subpackets.

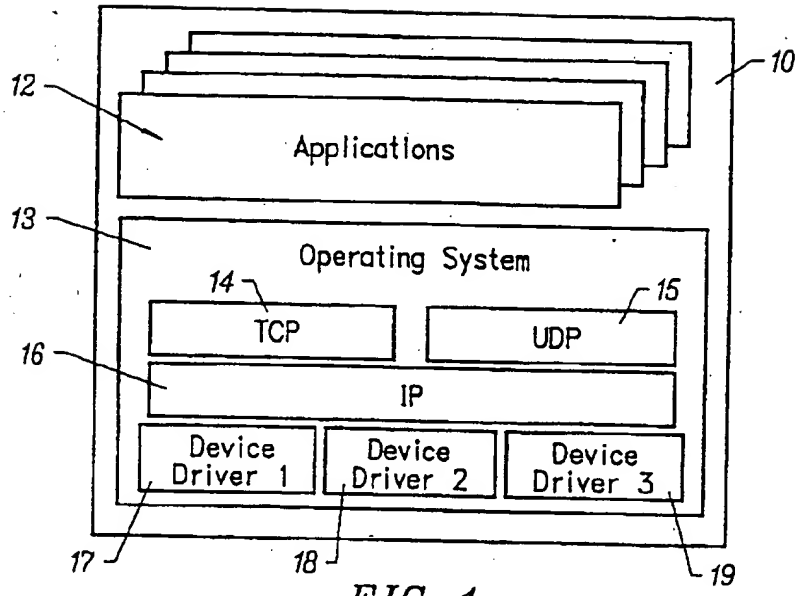


FIG. 1
PRIOR ART

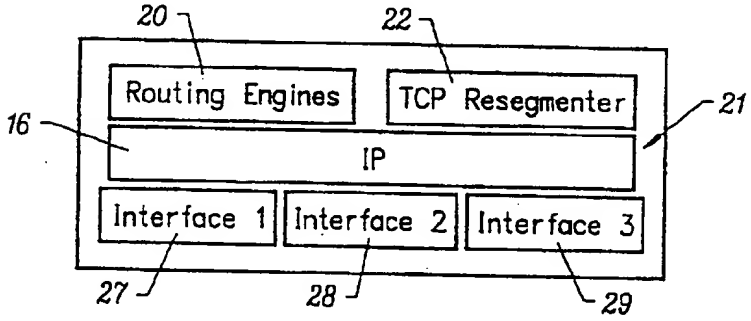


FIG. 2

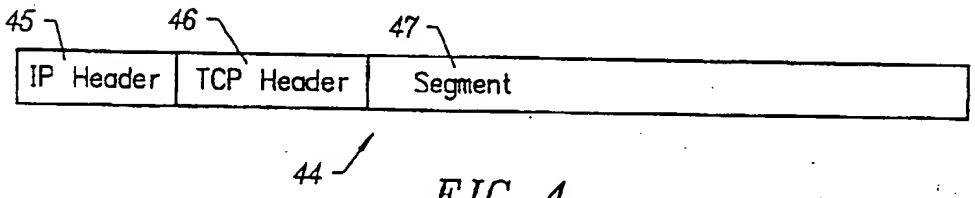


FIG. 4

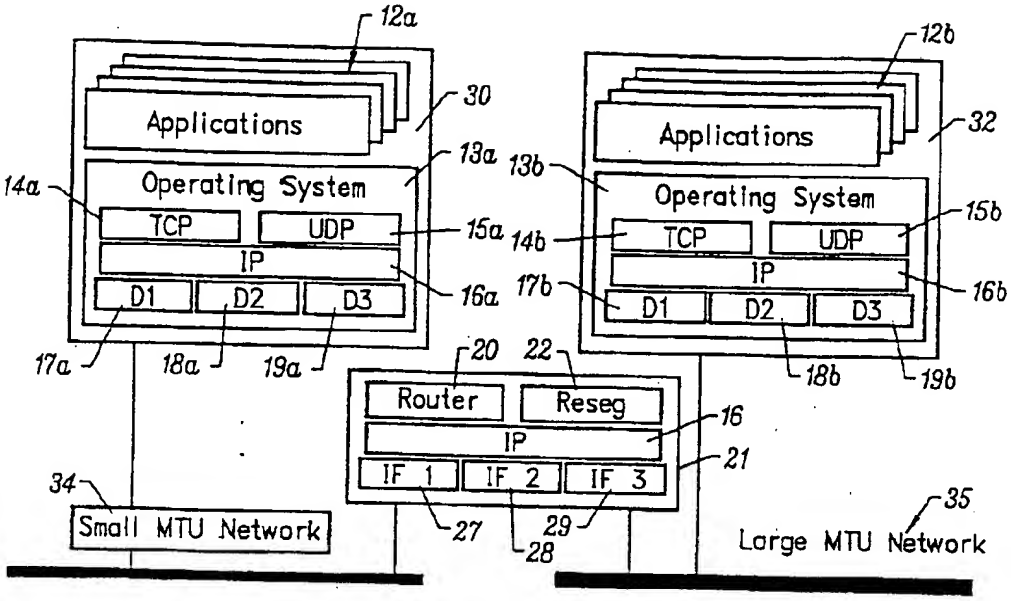

12a
12b
Applications
30
Applications
32
13a
13b
14a
Operating System
Operating System
15b
TCP
UDP
14b
TCP
UDP
15a
IP
IP
16b
16a
17b
D1
D2
D3
D1
D2
D3
17a
18a
19a
18b
19b
20
22
Router
Reseg
16
IP
21
IF 1
IF 2
IF 3
34
35
Small MTU Network
27
29
28
Large MTU Network

FIG. 3

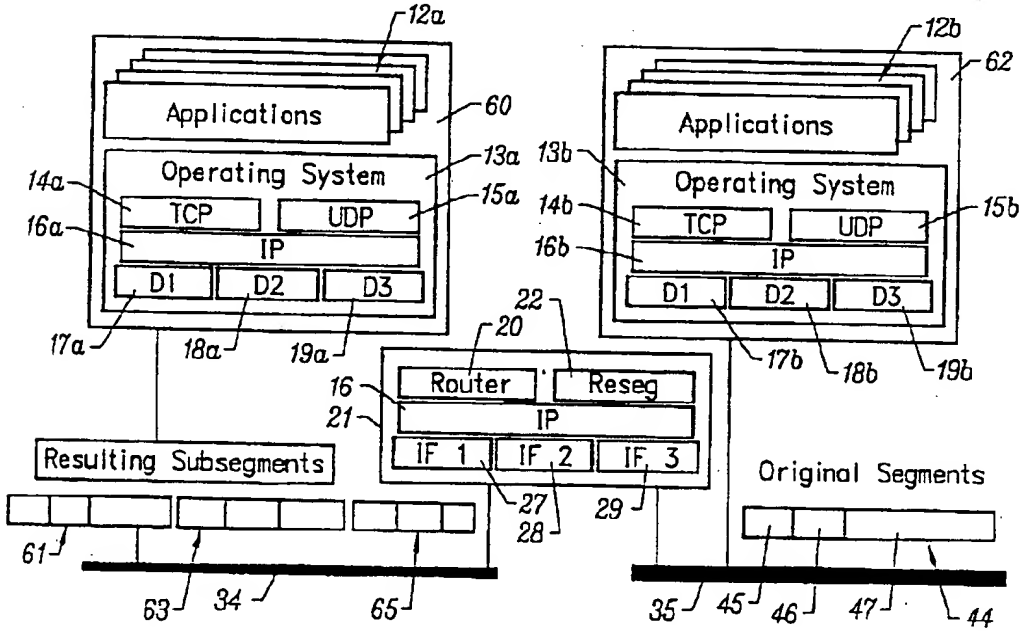

12a
12b
62
Applications
60
Applications
13a
13b
14a
Operating System
Operating System
15b
15a
TCP
UDP
14b
TCP
UDP
16a
IP
16b
IP
D1
D2
D3
D1
D2
D3
22
20
17a
18a
19a
17b
18b
19b
16
Router
Reseg
21
IP
IF 1
IF 2
IF 3
Resulting Subsegments
Original Segments
27
29
28
61
63
34
65
35
45
46
47
44

FIG. 6

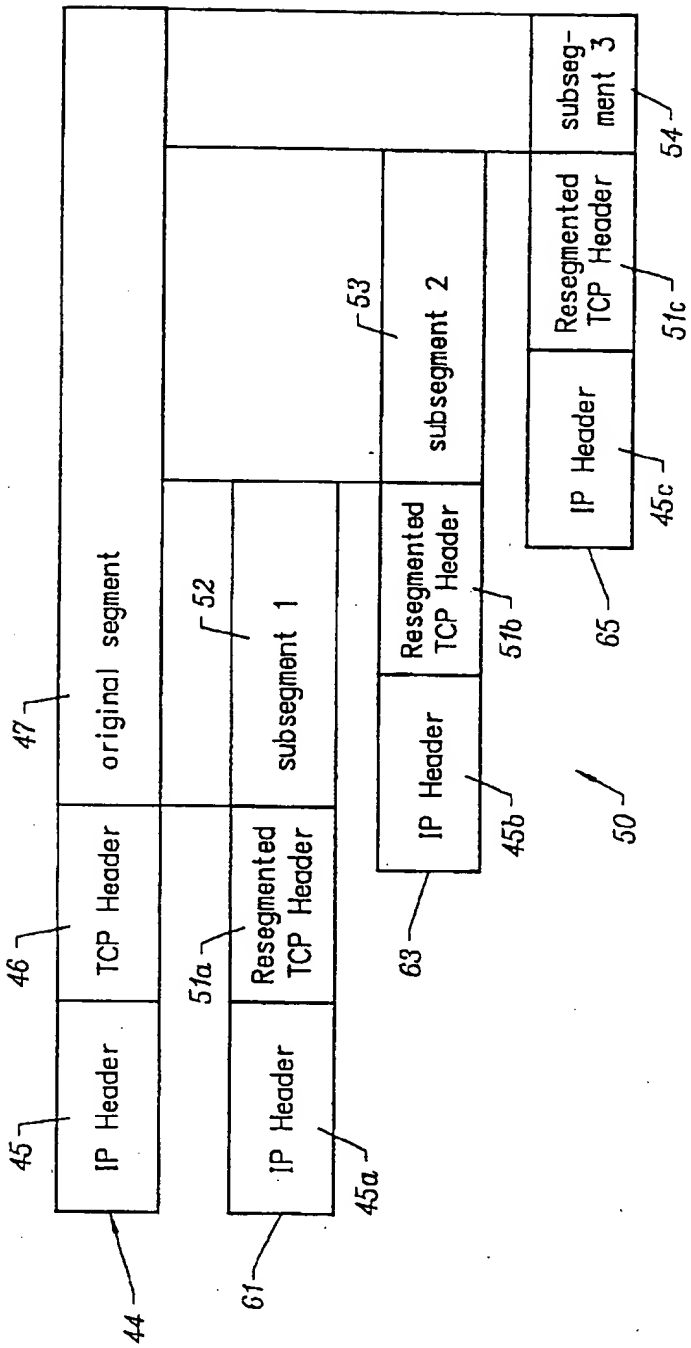

IP Header
TCP Header
original segment
IP Header
Resegmented TCP Header
subsegment 1
IP Header
Resegmented TCP Header
subsegment 2
IP Header
Resegmented TCP Header
subseg-ment 3
45
46
47
44
45a
51a
52
61
45b
51b
53
63
45c
51c
54
65
50

FIG. 5

ABSTRACT

A resegmentation entity implements a TCP resegmentation technique wherein a receiving host receives packets that appear as if it they have been transmitted specifically for the receiving host's MTU. The receiving host does not require the buffering and CPU utilization necessary for IP reassembly. Thus, the receiving host has a lower latency when receiving IP datagrams that contain resegmented TCP segments than it would if it needed to re-assemble an IP datagram from fragments before it could process the TCP segment. Further, the sending host transmits TCP segments at its largest MTU, without regard to the receiving station's MTU, knowing that the intermediate routing entity insures that TCP resegmentation occurs. In the event that an IP datagram containing a re-segmented TCP segment is lost, the sending host only has to retransmit the actual TCP data that was lost, and not the complete TCP segment.